RESEARCH ARTICLE / *ARTIGO*

Análise da epidemia da ferrugem do cafeeiro com árvore de decisão

Carlos A.A. Meira[1], Luiz H.A. Rodrigues[2] & Sérgio A. Moraes[3]

[1]Embrapa Informática Agropecuária, Cx. Postal 6041, 13083-970, Campinas, SP, Brasil; [2]Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Cx. Postal 6011, 13083-875, Campinas, SP, Brasil; [3]Centro de Pesquisa e Desenvolvimento de Fitossanidade, Instituto Agronômico de Campinas IAC, Cx. Postal 28, 13012-970, Campinas, SP, Brasil

Autor para correspondência: Carlos A.A. Meira, e-mail: carlos@cnptia.embrapa.br

RESUMO

Uma árvore de decisão foi desenvolvida com o objetivo de auxiliar na compreensão de manifestações epidêmicas da ferrugem do cafeeiro causada por *Hemileia vastatrix*. Taxas de infecção calculadas a partir de avaliações mensais de incidência da ferrugem foram agrupadas em três classes: redução ou estagnação - TX1; crescimento moderado (até 5p.p.) - TX2; e crescimento acelerado (acima de 5p.p.)- TX3. Dados meteorológicos, carga pendente de frutos do cafeeiro (*Coffea arabica*) e espaçamento entre plantas foram usados como variáveis explicativas das classes de taxa de infecção. A árvore de decisão foi treinada com 364 exemplos preparados a partir de dados coletados em lavouras de café em produção, de outubro de 1998 a outubro de 2006. Ela classificou corretamente 78% do conjunto de treinamento e a sua acurácia foi estimada em 73% para a classificação de novos exemplos. O acerto do modelo foi de 88%, 57% e 79% dos exemplos, respectivamente, para as classes de taxa de infecção TX1, TX2 e TX3. As variáveis explicativas mais importantes foram a temperatura média nos períodos de molhamento foliar, a carga pendente de frutos, a média das temperaturas máximas diárias no período de incubação e a umidade relativa do ar. A árvore de decisão demonstrou seu potencial como modelo de representação simbólica e interpretável, permitindo a identificação das fronteiras de decisão existentes nos dados e da lógica contida neles, auxiliando na compreensão de quais variáveis e como as interações dessas variáveis conduziram as epidemias da ferrugem do cafeeiro no campo.
Palavras-chave: *Hemileia vastatrix*, *Coffea arabica*, descoberta de conhecimento em bases de dados, mineração de dados.

ABSTRACT

Analysis of coffee leaf rust epidemics with decision tree

A decision tree was developed to aid the understanding of coffee rust epidemics caused by *Hemileia vastatrix*. Infection rates calculated from monthly assessments of rust incidence were grouped into three classes: reduction or stagnation - TX1; moderate growth (up to 5pp) - TX2; and accelerated growth (above 5pp) - TX3. Meteorological data, expected yield and space between plants were used as explanatory variables for the infection rate classes. The decision tree was trained using 364 examples prepared from data collected in coffee-growing areas between October 1998 and October 2006. The model correctly classified 78% of the training data set and its accuracy was estimated at 73% for the classification of new examples. The success rates of the model were 88%, 57% and 79%, respectively, for the infection rate classes TX1, TX2 and TX3. The most important explanatory variables were mean temperature during leaf wetness periods, expected yield, mean of maximum temperatures during the incubation period and relative air humidity. The decision tree demonstrated its potential as a symbolic and interpretable model. Its model representation identified the existing decision boundaries in the data and the logic underlying them, helping to understand which variables, and interactions between these variables, led to coffee rust epidemics in the field.
Keywords: *Hemileia vastatrix*, *Coffea arabica*, knowledge discovery in databases – KDD, data mining.

INTRODUÇÃO

A ferrugem, cujo agente etiológico é *Hemileia vastatrix* Berk. & Br., é a principal doença do cafeeiro (*Coffea arabica* L.) em todo o mundo. No Brasil, os prejuízos na produção atingem cerca de 35% nas regiões onde as condições climáticas são favoráveis à doença (Zambolim *et al.*, 1997). O conhecimento dos fatores que determinam a maior taxa de progresso da ferrugem é de grande importância, uma vez que eles condicionam a distribuição da doença, a sua incidência e a severidade. O estudo das relações entre o patógeno, o hospedeiro e o ambiente pode auxiliar na compreensão da ocorrência de epidemias e, consequentemente, permitir a aplicação de medidas de controle mais adequadas (Montoya & Chaves, 1974).

A epidemiologia da ferrugem do cafeeiro já foi tema

Parte da Tese de Doutorado do primeiro autor. Universidade Estadual de Campinas. Campinas SP.

de diversos trabalhos. A maioria desses estudos utilizou a regressão múltipla para ajustar os dados (Kushalappa *et al.*, 1983; Kushalappa & Eskes, 1989; Montoya & Chaves, 1974; Moraes *et al.*, 1976; Zambolim *et al.*, 2002). Estudos mais recentes procuraram empregar outras técnicas, como a análise de trilha (Silva-Acuña *et al.*, 1998) e as redes neurais (Pinto *et al.*, 2002). Outra técnica alternativa, ainda pouco utilizada em epidemiologia de doenças de plantas, é a indução de árvores de decisão. As árvores de decisão são de interesse especial para a mineração de dados ou descoberta de conhecimento em bases de dados (Fayyad *et al.*, 1996), pois utilizam representações simbólicas e interpretáveis. Soluções simbólicas permitem a compreensão das fronteiras de decisão que existem nos dados e também da lógica implícita neles (Apte & Weiss, 1997). Redes neurais, por exemplo, embora uma ferramenta preditiva poderosa, são relativamente difíceis de compreender quando comparadas com as árvores de decisão (Fayyad *et al.*, 1996). Multicolinearidade entre as variáveis independentes não afeta o desempenho das árvores de decisão, diferentemente das técnicas de regressão (Butt & Royle, 1974; Hand *et al.*, 2001). Diversas variáveis, numéricas ou categóricas, podem ser analisadas ao mesmo tempo, sendo que o próprio algoritmo de indução se encarrega de selecionar as de maior importância.

A árvore de decisão é um modelo representado graficamente por nós e ramos, parecido com uma árvore, mas no sentido invertido (Han & Kamber, 2001; Monard & Baranauskas, 2002a; Witten & Frank, 2005). O nó raiz é o primeiro nó da árvore, no topo da estrutura. Os nós internos, incluindo o nó raiz, são nós de decisão. Cada um contém um teste sobre uma variável independente e os resultados desse teste formam os ramos da árvore. Os nós folhas, nas extremidades da árvore, representam valores de predição para a variável dependente ou distribuições de probabilidade desses valores. As árvores de decisão são também chamadas de árvores de classificação ou de regressão, caso a variável dependente seja categórica ou numérica, respectivamente. O propósito básico da indução de uma árvore de decisão é produzir um modelo de predição preciso ou descobrir a estrutura preditiva do problema (Breiman *et al.*, 1984). No último caso, a intenção é compreender quais variáveis e interações dessas variáveis conduzem o fenômeno estudado. Esses dois propósitos não são excludentes, podendo aparecer juntos em um mesmo estudo.

Algumas experiências têm sido relatadas na literatura com respeito à aplicação de árvores de decisão na fitopatologia. Paul & Munkvold (2004) usaram este tipo de modelagem para avaliação de risco da cercosporiose do milho (*Cercospora zeae-maydis* Tehon & E.Y. Daniels). Árvores de decisão e modelos de regressão logística foram usados para predizer a severidade da doença em estágio final do cultivo, a partir de dados obtidos no pré-plantio e de características do genótipo do milho (*Zea mays* L.). Classes (categorias) de severidade da cercosporiose foram definidas e serviram como os valores da variável resposta. A abordagem de modelagem CART, de "Classification and Regression Trees" (Breiman *et al.*, 1984), e diferentes abordagens de regressão logística foram empregadas para predizer as classes de severidade em função da data do plantio, da quantidade de resíduo de milho no solo, da seqüência de plantio, da maturidade do genótipo e do seu nível de resistência à cercosporiose do milho e da longitude. Foram usados 332 casos para desenvolvimento dos modelos e 30 casos independentes para validação. Os modelos de regressão logística classificaram corretamente de 60 a 70% dos casos de validação, enquanto as árvores de decisão classificaram corretamente de 57 a 77% desses mesmos casos. Ambas as abordagens mostraram potencial como ferramentas de tomada de decisão no gerenciamento da doença.

Árvores de decisão foram usadas por Molineros *et al.* (2004; 2005) para modelar epidemias de giberela do trigo [*Gibberella zeae* (Schwein.) Petch]. A doença foi codificada como uma variável binária, com valor 1 atribuído aos casos com severidade maior ou igual a 10% e valor 0, caso contrário. Cada caso consistiu de variáveis meteorológicas horárias, incluindo temperatura, umidade relativa e chuva, sintetizadas para sete dias antes da data de florescimento da cultura. Um total de 154 casos foram usados, 70% para modelagem (108 casos) e os restantes 30% para validação (46 casos). O relacionamento entre as condições do tempo e a doença foi modelado com regressão logística, redes neurais, K-vizinhos mais próximos e árvores de decisão. Os modelos foram avaliados quanto à habilidade de classificar corretamente os casos, bem como quanto à sensitividade (capacidade de classificar corretamente casos com severidade $\geq$ 10%) e especificidade (capacidade de classificar corretamente casos com severidade $<$ 10%). As árvores de classificação e os modelos de regressão logística tiveram os melhores desempenhos.

A indução de árvores de decisão tem sido usada também para se obter modelos de estimativa de uma importante variável na epidemiologia de muitas doenças, a duração do período de molhamento foliar. Gleason *et al.* (1994) desenvolveram um modelo empírico não paramétrico para a estimativa da duração do molhamento foliar devido ao orvalho, usando árvores de classificação e regressão (CART) em conjunto com análise discriminante linear. Posteriormente, esse modelo foi usado para melhorar a estimativa de duração do período de molhamento foliar em comparação com um modelo proprietário (Kim *et al.*, 2002).

Consideradas as vantagens mencionadas das árvores de decisão em relação às outras técnicas de modelagem e verificadas algumas de suas aplicações na fitopatologia, o objetivo deste trabalho foi aplicar e avaliar o potencial da técnica de indução de árvores de decisão na epidemiologia da ferrugem do cafeeiro. O intuito foi obter uma árvore de decisão capaz de auxiliar na compreensão de como as condições do ambiente, a carga pendente de frutos do cafeeiro e o espaçamento entre as plantas na lavoura condicionaram a taxa de infecção da doença, identificando, dentre estes,

os fatores mais importantes no progresso da ferrugem do cafeeiro no campo.

MATERIAL E MÉTODOS

Os dados utilizados foram coletados por Japiassú *et al.* (2007) e se referem ao acompanhamento mensal da incidência da ferrugem do cafeeiro, de outubro de 1998 a outubro de 2006, na fazenda experimental da Fundação Procafé, localizada em Varginha, MG, latitude sul de 21º 34' 00'', longitude oeste de 45º 24' 22'' e altitude de 940 m. A cada ano, no mês de setembro, foram selecionadas oito lavouras de café em produção, quatro em espaçamento largo (por volta de 3,5 m entre linhas e 0,7 m entre plantas – densidade média de 4.000 plantas/ha) e quatro adensadas (por volta de 2,5 m entre linhas e 0,5 m entre plantas – densidade média de 8.000 plantas/ha). Para cada espaçamento, foram escolhidas duas lavouras com alta carga pendente de frutos (acima de 30 sacas beneficiadas/ha) e duas com baixa carga (abaixo de 10 sacas beneficiadas/ha). Em cada par de lavouras, uma era da cultivar Catuaí e a outra da cultivar Mundo Novo. Não houve controle da doença durante o ano agrícola nos talhões escolhidos.

O processo de amostragem, realizado no final de cada mês, foi o recomendado por Chalfoun (1997): coleta de 100 folhas do terço médio das plantas em cada talhão, entre o terceiro e o quarto par de folhas; contagem do número de folhas com lesões de ferrugem; e determinação da incidência (percentual de folhas atacadas) para cada uma das quatro combinações de espaçamento e produção das lavouras. Dados meteorológicos, como temperatura (média, máxima e mínima), precipitação pluviométrica, umidade relativa do ar, entre outros, foram registrados a cada 30 min por uma estação meteorológica automática (marca Davis, modelo Groweather Industrial) instalada próximo dos locais de avaliação da incidência da ferrugem.

A análise dos dados foi conduzida como um processo de descoberta de conhecimento em bases de dados (Fayyad *et al.*, 1996), de acordo com o modelo de processo de mineração de dados CRISP-DM (Chapman *et al.*, 2000). O processo compreendeu as fases de compreensão do domínio, de entendimento dos dados, de preparação dos dados, de modelagem e de avaliação. As fases principais estão descritas a seguir.

Preparação dos dados

A variável dependente (atributo classe) foi obtida por meio de transformações nos valores de incidência da ferrugem do cafeeiro. A partir dos valores de incidência, foram calculadas as taxas de infecção de cada mês, subtraindo-se a incidência do mês com a do mês anterior. Em seguida, o valor numérico da taxa de infecção foi mapeado para uma classe ou categoria: 'TX1(<=0)', para taxas de infecção negativas ou nulas; 'TX2(>0<=5)', para taxas de infecção positivas, menores ou iguais a 5 pontos percentuais (p.p.) e 'TX3(>5)', para taxas de infecção maiores que 5p.p. Estas classes foram escolhidas com base nas faixas de valores de incidência da ferrugem do cafeeiro recomendadas por Zambolim *et al.* (1997) para o controle da doença via foliar.

As variáveis explicativas (independentes) meteorológicas foram construídas a partir do nível horário (registros da estação), passando pelo nível diário, até um nível que permitisse a análise de seu relacionamento com a variável dependente. No nível diário, além de médias e de somatórios das variáveis meteorológicas, foram calculados valores estimados de molhamento foliar prolongado (mínimo de 6 h), uma vez que a germinação só ocorre se a folha estiver molhada, e seis horas de água livre na superfície da folha foi avaliado como o tempo mínimo necessário para ocorrer infecção (Kushalappa *et al.*, 1983). O número de horas com alta umidade relativa do ar (≥ 95%) foi utilizado como medida indireta de molhamento foliar contínuo (Sutton *et al.*, 1984). Em dias com períodos de molhamento disjuntos, foi considerado o maior, com tolerância de até uma hora entre eles para juntar em um único período. Os períodos de molhamento foliar foram analisados tanto na sua extensão total como na sua fração noturna (das 20:00 h às 8:00 h), já que a infecção ocorre preferencialmente na ausência de ou com pouca luminosidade (Montoya & Chaves, 1974). O dia foi considerado de 12:00 h de um dia comum até 12:00 h do dia seguinte, pois os períodos de molhamento ocorrem geralmente entre um dia e o outro. As temperaturas médias durante os períodos total e noturno de molhamento foliar contínuo também foram calculadas para cada dia, uma vez que, enquanto a superfície da folha está molhada, a temperatura é o fator principal que determina o percentual de germinação dos esporos e de penetração (Kushalappa *et al.*, 1983).

Na seqüência da preparação dos dados meteorológicos, cada dia foi tratado como um potencial dia de infecção. Considerando um período de incubação estimado, cada dia foi associado ao mês correspondente de avaliação da incidência da ferrugem (Figura 1). O período de incubação para cada dia foi estimado pela equação proposta por Moraes *et al.* (1976):

$$y = 103{,}01 - 0{,}98x_1 - 2{,}1x_2$$

onde y corresponde à estimativa do período de incubação em dias; x_1 à temperatura média máxima e x_2 à temperatura média mínima durante o período. Dessa forma, cada dia foi associado a uma taxa de infecção, para a qual possivelmente teve parcela de contribuição. O conjunto de dias associado a uma taxa de infecção foi denominado de período de infecção (PINF) (Figura 1). As variáveis explicativas meteorológicas usadas na modelagem foram derivadas para cada um desses períodos de infecção (Tabela 1). O espaçamento da lavoura (lavoura adensada ou larga) e a carga pendente de frutos (carga alta ou baixa) completaram o conjunto das variáveis explicativas (Tabela 1).

A preparação dos dados foi feita, em grande parte, por meio de programas de computador escritos na linguagem de programação Perl (ActivePerl© versão 5.8.7, ActiveState

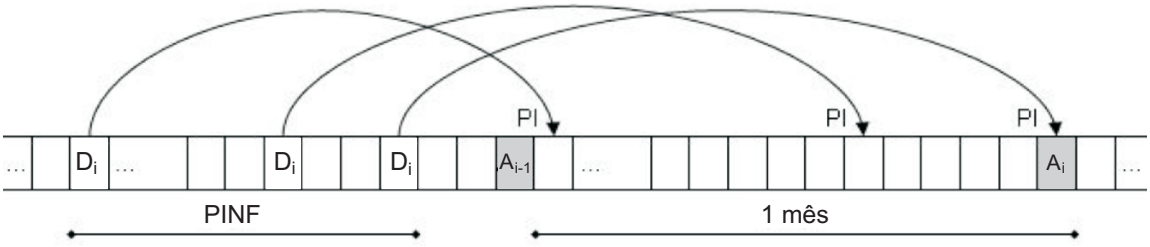

FIG. 1: Representação dia-a-dia do esquema usado na preparação dos dados meteorológicos. D_i – dia de infecção, A_i – avaliação da incidência da ferrugem do cafeeiro, A_{i-1} – avaliação da incidência no mês anterior, PI – período de incubação e PINF – período de infecção.

TABELA 1 - Relação das variáveis usadas na indução da árvore de decisão

	Unidade de Descrição		
Nome	**Tipo**	**Medida**	**Significado**
Var. dependente			
TAXA_INF3N	nominal	-	Taxa de infecção em três níveis categóricos: TX1(<=0), TX2(>0<=5) ou TX3(>5).
Var. explicativas			
CARGA	binário	-	Carga pendente de frutos: ALTA ou BAIXA.
LAVOURA	binário	-	Espaçamento: lavoura ADENSADA ou LARGA.
MED_PRECIP_PINF	numérico	mm	Média das precipitações pluviais diárias no PINF (período de infecção).
NHNUR95_PINF	numérico	h	Média diária do número de horas noturnas com umidade relativa do ar ≥ 95%
NHUR95_PINF	numérico	h	Média diária do número de horas com umidade relativado ar ≥ 95% no PINF.
PRECIP_PINF	numérico	mm	Precipitação pluvial acumulada no PINF.
THNUR95_PINF	numérico	ºC	Temperatura média diária durante as horas noturnas com umidade relativa do ar ≥ 95% no PINF.
THUR95_PINF	numérico	ºC	Temperatura média diária durante as horas com umidade relativa do ar ≥ 95% no PINF.
TMAX_PINF	numérico	ºC	Média das temperaturas máximas diárias no PINF.
TMAX_PI_PINF	numérico	ºC	Média das temperaturas máximas diárias no período de incubação para os dias do PINF.
TMED_PINF	numérico	ºC	Média das temperaturas médias diárias no PINF.
TMIN_PINF	numérico	ºC	Média das temperaturas mínimas diárias no PINF.
TMIN_PI_PINF	numérico	ºC	Média das temperaturas mínimas diárias no período de incubação para os dias do PINF.
UR_PINF	numérico	%	Umidade relativa do ar média diária no PINF.

Corp.). Alguns procedimentos finais foram realizados com o "software" SAS® Enterprise Miner™ (versão 4.3, SAS Institute Inc.). O conjunto de dados preparado totalizou 384 exemplos ou casos (8 anos x 12 meses x 4 combinações espaçamento-carga). No entanto, houve períodos de falha no registro da estação meteorológica, por falta de energia, funcionamento incorreto ou manutenção nos meses de novembro e dezembro de 2000 e fevereiro de 2000 e 2003. Os dados meteorológicos disponíveis também não foram suficientes para formar o período de infecção correspondente à incidência observada em outubro de 1998. Sendo assim, os exemplos dos referidos meses (5 meses x 4 combinações = 20) foram eliminados, resultando no conjunto de modelagem com 364 exemplos, chamado de conjunto de treinamento.

Modelagem

O algoritmo básico de indução de árvores de decisão constrói a árvore de forma recursiva, de cima para baixo (Han & Kamber, 2001). Inicia com o conjunto de treinamento, que é dividido de acordo com um teste sobre uma das variáveis independentes, formando-se subconjuntos mais homogêneos em relação à variável dependente. Esse procedimento é repetido até que se consiga conjuntos de exemplos bem homogêneos, para os quais seja possível atribuir um único valor para a variável dependente. O

critério utilizado para escolher a variável independente que divide o conjunto de exemplos em cada repetição é o aspecto principal do processo de indução. Dentre os critérios mais conhecidos e usados, cita-se o índice Gini (Breiman *et al.*, 1984) e o ganho de informação (Quinlan, 1993), relacionado com a redução da entropia dos exemplos, que foi o critério escolhido para este trabalho.

A indução da árvore de decisão foi realizada por meio da ferramenta "Tree" ("Tree node") do SAS® Enterprise Miner™. A árvore de decisão foi escolhida para ser binária, com dois ramos a partir de cada nó interno. Para evitar que o modelo ficasse muito específico para o conjunto de treinamento ("overfitting"), o que comprometeria a sua generalização e o desempenho com novos exemplos, foram adotadas duas regras de parada do algoritmo de indução. A primeira regra limitou a profundidade da árvore, permitindo-a ter no máximo seis níveis (o nó raiz é considerado estar no nível zero). A segunda regra limitou a fragmentação do conjunto de treinamento, requerendo um mínimo de dez exemplos em cada nó para a busca de uma nova divisão e pelo menos cinco exemplos em cada nó folha. Além das regras de parada, denominadas de pré-poda, foi realizado um procedimento de pós-poda, após a indução da árvore completa. Junto com essa árvore completa, foram avaliadas todas as suas possíveis sub-árvores e escolhida a menor sub-árvore (menor complexidade) com a menor taxa de erro sobre o conjunto de treinamento.

Avaliação do modelo

A taxa de erro e a acurácia, que é o complemento da taxa de erro, são as medidas de avaliação mais comuns para árvores de decisão (Han & Kamber, 2001; Witten & Frank, 2005). A taxa de erro é a proporção de erros de predição sobre um conjunto de exemplos em que se conhece o valor da variável dependente. Um conjunto de exemplos pode ser denotado por $\{(x_i, y_i), i = 1, 2, ..., n\}$, onde x_i é um vetor de valores das variáveis independentes, y_i é o valor da variável dependente e n é a quantidade de exemplos. A taxa de erro de uma árvore de decisão h sobre este conjunto de exemplos é dada pela equação:

$$err(h) = \frac{1}{n}\sum_{i=1}^{n} D\left(h(x_i), y_i\right)$$

onde $h(x_i)$ é o valor de predição para x_i e $D(a, b)$ é igual a 1, se a é diferente de b, ou igual a 0, caso contrário. A partir da taxa de erro, a acurácia é dada pela equação:

$$acc(h) = 1 - err(h)$$

Calcular a taxa de erro sobre o conjunto de treinamento (método de resubstituição) normalmente resulta em uma estimativa altamente otimista, devido à especialização do modelo com respeito aos exemplos. Uma das formas mais usadas de contornar esse problema é dividir aleatoriamente os exemplos em dois conjuntos independentes, um de treinamento e o outro de validação (Han & Kamber, 2001). O conjunto de treinamento é usado para induzir o modelo e a sua taxa de erro é estimada a partir do conjunto de validação. Outro método bastante usado é a validação cruzada ("cross-validation"), particularmente quando a quantidade de dados para dividir entre treinamento e validação é limitada (Witten & Frank, 2005). Na validação cruzada, os exemplos são aleatoriamente divididos em k partições mutuamente exclusivas ("folds") de tamanho aproximadamente igual. Uma das partições é reservada para validação, enquanto as demais juntas são usadas para treinamento. Este procedimento é executado k vezes, cada vez com uma partição diferente para a validação. A taxa de erro é calculada como a média das taxas de erro obtidas em cada uma das partições de validação. A vantagem da validação cruzada é usar cada um dos exemplos tanto para treinamento quanto para validação.

No presente trabalho, a taxa de erro e a acurácia foram estimadas pelos métodos de resubstituição e de validação cruzada com 10 partições aleatórias do conjunto de treinamento ("10-fold cross-validation"). Testes extensivos em muitos e diferentes conjuntos de dados mostraram que dez é um número próximo do número exato de partições para se obter a melhor estimativa de erro (Witten & Frank, 2005).

Além dessas medidas, foi produzida a matriz de confusão da árvore de decisão, a qual oferece meios efetivos para a avaliação do modelo com base em cada classe (Monard & Baranauskas, 2002b). Cada elemento da matriz mostra o número de exemplos para os quais a classe verdadeira é a linha e a classe predita é a coluna. A diagonal principal da matriz (elementos (i,j), onde $i = j$) representa os acertos do modelo, enquanto os demais elementos representam os erros, discriminados para cada classe. Cada elemento da matriz (M) foi calculado segundo a equação:

$$M(C_i, C_j) = \sum_{\{\forall (x,y) \in T; y = C_i\}} I\left(h(x), C_j\right)$$

onde $i, j = 1, 2, ..., k$; $\{C_1, C_2, ..., C_k\}$ é o conjunto das classes para a variável dependente; (x, y) são os exemplos do conjunto de treinamento T e $I(a, b)$ é igual a 1, se a é igual a b, ou igual a 0, caso contrário.

RESULTADOS

O início da epidemia da ferrugem do cafeeiro, na média de todos os anos, foi no mês de dezembro e atingiu o pico no mês de junho, independente da combinação de espaçamento e de carga pendente de frutos da lavoura (Figura 2). A partir de dezembro, as taxas de infecção atingiram níveis mais elevados, com o percentual de distribuição da classe de taxa de infecção 'TX3(>5)', para taxas de infecção maiores que 5p.p., ultrapassando o das classes de menor nível (Figura 3). O percentual de distribuição das três classes de taxa de infecção da ferrugem do cafeeiro, no

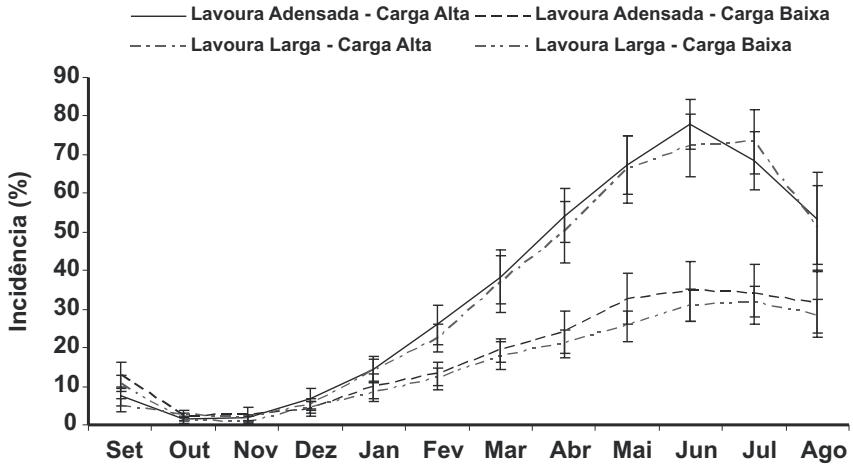

FIG. 2 - Evolução mensal da incidência da ferrugem do cafeeiro (% de folhas com lesões) em lavouras com diferentes espaçamentos e cargas pendentes de frutos – média de 1998/1999 a 2005/2006 na Fazenda Experimental de Varginha. Varginha, MG. A barra indica o erro (desvio padrão da média). Fundação Procafé. Varginha MG.

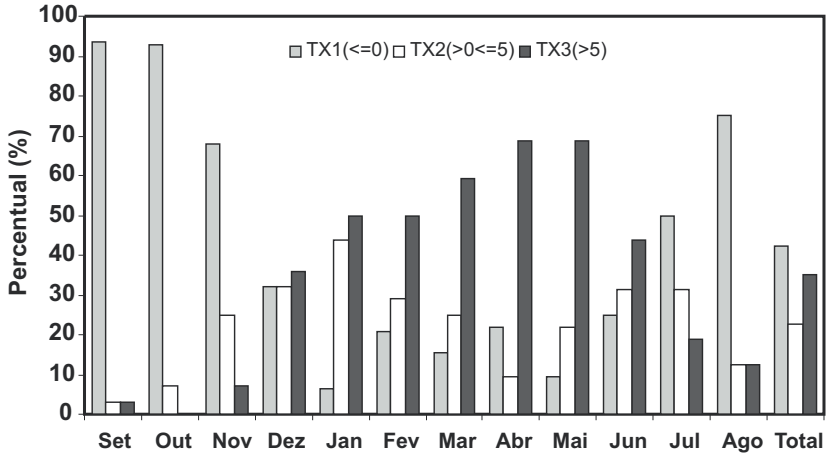

FIG. 3 - Distribuição percentual das três classes de taxa de infecção da ferrugem do cafeeiro, para cada mês e para o total dos meses, referente ao período de 1998/1999 a 2005/2006 na Fazenda Experimental de Varginha, Fundação Procafé. Varginha MG.

conjunto de todos os meses do período analisado, foi 42% (154 exemplos) da classe 'TX1(<=0)', 23% (82 exemplos) da classe 'TX2(>0<=5)' e 35% (128 exemplos) da classe 'TX3(>5)' (Figura 3).

A árvore de decisão gerada ajuda a compreender o relacionamento entre as variáveis explicativas e a taxa de infecção da ferrugem do cafeeiro (Figura 4). Deve-se partir do topo da árvore (nó "raiz") e descer pelos seus ramos, de acordo com os testes nas variáveis explicativas, até se chegar nos nós "folhas". A primeira variável usada na decisão da classe da taxa de infecção (Figura 4, nó 1) foi a temperatura média nos períodos de alta umidade relativa do ar (THUR95_PINF). Temperaturas inferiores a 17ºC produziram taxas de infecção negativas ou nulas na maioria dos casos (73%), enquanto temperaturas maiores ou iguais a 17ºC resultaram em taxas de infecção positivas na maior parte das vezes (28% de 'TX2(>0<=5)' e 57% de 'TX3(>5)').

O próximo teste, descendo pelo ramo à esquerda do nó "raiz" (Figura 4, nó 2,), foi escolhido sobre a média das temperaturas máximas diárias no período de incubação (TMAX_PI_PINF). Temperaturas maiores ou iguais a 25,95ºC resultaram em taxas de infecção negativas ou nulas (93% dos casos). Neste ponto chega-se a um nó folha (Figura 4, nó 5) e, caso se estivesse classificando novos exemplos, para os quais não se conhece o valor da variável dependente, a árvore de decisão indicaria a taxa de infecção provável desses exemplos como da classe 'TX1(<=0)'. O caminho de decisão entre o nó raiz e o nó folha pode ser traduzido para uma regra na forma 'SE <*condição*> ENTÃO <*decisão*>'. Assim, o caminho até o nó 5 se traduz na regra 'SE (THUR95_PINF < 17) e (TMAX_PI_PINF >= 25,95) ENTÃO TAXA_INF3N = TX1(<=0)'.

Em relação ainda a THUR95_PINF, a árvore de decisão indica que temperaturas abaixo de 16°C foram desfavoráveis à infecção (Figura 4, nó 8) e que temperaturas maiores ou iguais a 19,35°C foram bastante

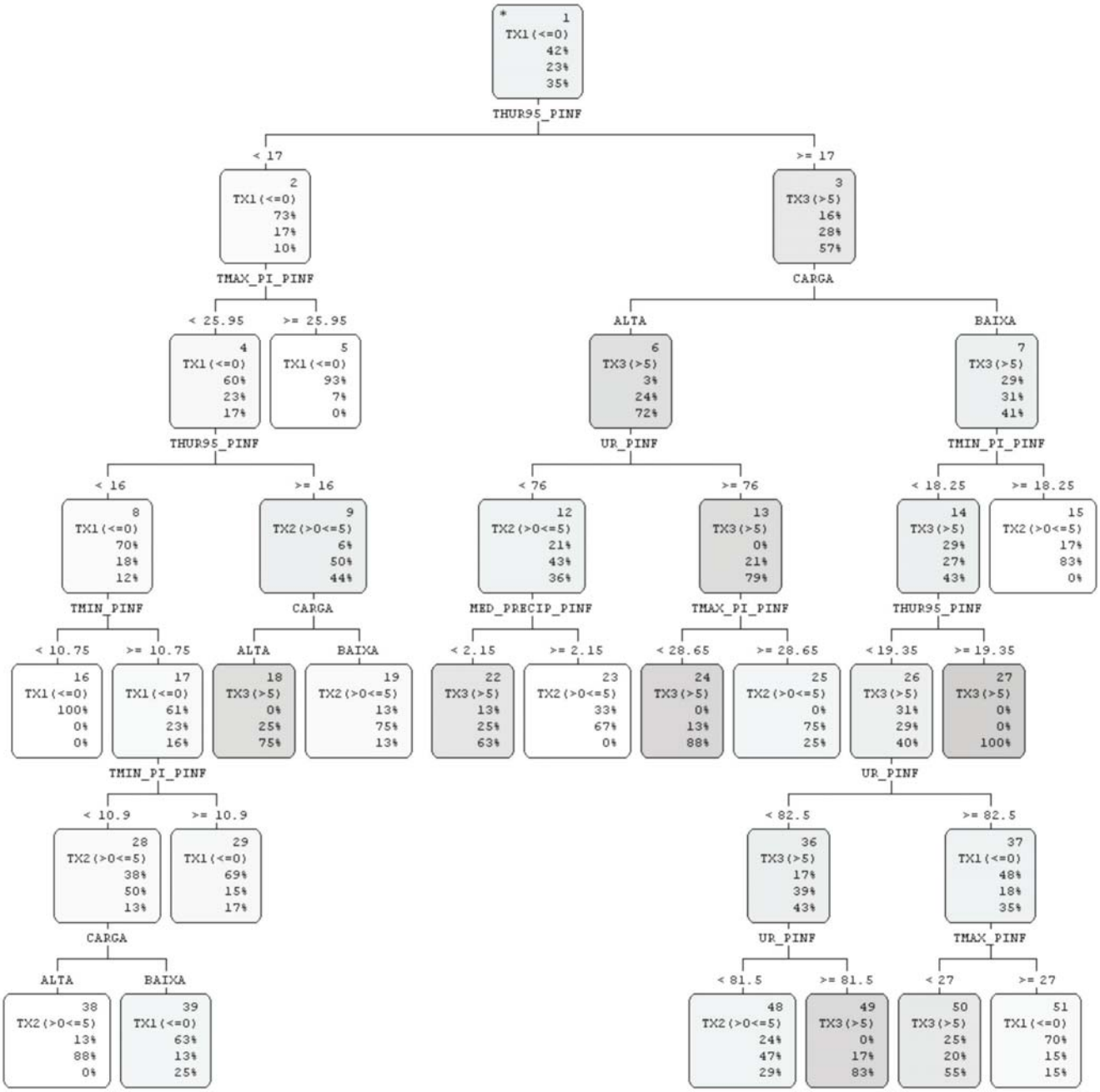

FIG. 4 - Árvore de decisão que auxilia na compreensão das epidemias de ferrugem do cafeeiro, de acordo com classes de taxa de infecção, em função de variáveis meteorológicas e da carga pendente de frutos. *As informações em cada nó representam, de cima para baixo, o número identificador do nó, a classe de taxa de infecção predominante e a distribuição percentual de cada classe no nó, na ordem 'TX1(<=0)', 'TX2(>0<=5)' e 'TX3(>5)'. Os nós são coloridos em tons de cinza com base na proporção de exemplos da classe 'TX3(>5)', de 0% (descolorido) a 100% (cinza escuro).

favoráveis à infecção (Figura 4, nó 27).

Segundo as decisões, com base na carga pendente de frutos (Figura 4, nós 3, 9 e 28), taxas de infecção em níveis mais elevados ocorreram em cafeeiros com carga pendente alta, em comparação com os de carga pendente baixa.

Médias elevadas de temperatura máxima diária no período de incubação (TMAX_PI_PINF) tiveram efeito depressivo sobre a taxa de infecção (Figura 4, nó

25), parecido com o efeito encontrado no nó 5, apesar da diferença dos limiares de decisão e das proporções de cada classe nos dois nós. Médias altas de temperatura mínima diária no período de incubação (TMIN_PI_PINF) também tiveram efeito depressivo sobre a taxa de infecção (Figura 4, nó 15). O teste sobre TMIN_PI_PINF no nó 17 parece não ter estabelecido um limite entre uma condição favorável e outra menos favorável, sendo que a decisão final da classe da taxa de infecção dependeu ainda de outras variáveis.

Valores médios diários mais elevados de umidade relativa do ar (UR_PINF) corresponderam a níveis mais elevados da taxa de infecção (nós 13 e 49, Figura 4). A decisão com base em UR_PINF no nó 26 parece também não ter dividido entre uma condição favorável e outra desfavorável. Médias baixas de temperatura mínima diária (TMIN_PINF) exerceram influência negativa nas taxas de infecção (Figura 4, nó 16). Valores médios elevados de temperatura máxima diária (TMAX_PINF) também exerceram influência negativa nas taxas de infecção (Figura 4, nó 51).

A precipitação pluvial média diária (MED_PRECIP_PINF) foi escolhida no teste para o nó 12 (Figura 4), com efeito negativo nas taxas de infecção para valores maiores ou iguais a 2,15 mm (Figura 4, nó 23). A precipitação acumulada no período de infecção (PRECIP_PINF) poderia ter sido escolhida no teste para o mesmo nó 12. Nesse caso, o limiar de decisão, em vez de 2,15 mm para MED_PRECIP_PINF, seria 73 mm para PRECIP_PINF.

A árvore de decisão apresentou acurácia de 78% sobre o conjunto de treinamento e a acurácia obtida utilizando validação cruzada ("10-fold cross-validation") foi de 73% (Tabela 2). Em relação aos acertos para cada classe de taxa de infecção, 88% (135 exemplos) da classe 'TX1(<=0)', 57% (47 exemplos) da classe 'TX2(>0<=5)' e 79% (101 exemplos) da classe 'TX3(>5)' foram corretamente classificados (Tabela 3). Quanto aos erros, por exemplo, 20% (16 exemplos) da classe 'TX2(>0<=5)' foram classificados como da classe 'TX1(<=0)' e 23% (19 exemplos) classificados como da classe 'TX3(>5)' (Tabela 3).

DISCUSSÃO

A importância da temperatura durante o período de molhamento foliar no progresso da ferrugem do cafeeiro é reconhecida na literatura (Kushalappa & Eskes, 1989; Moraes, 1983; Zambolim *et al.*, 1997; Zambolim *et al.*, 2002). Enquanto a superfície da folha está molhada, a temperatura é o fator principal que determina o percentual de germinação dos esporos e de penetração do agente etiológico da ferrugem (Kushalappa *et al.*, 1983). Na árvore de decisão gerada, a temperatura durante o molhamento foliar, medida indiretamente pela temperatura média nos períodos de alta umidade relativa do ar (THUR95_PINF), foi a variável mais importante na determinação da classe de taxa de infecção da ferrugem. Foi escolhida para o primeiro teste no nó raiz e para outros dois testes nos níveis intermediários da árvore de decisão.

THUR95_PINF inferiores a 17 e 16ºC cada vez mais desfavoráveis à infecção e superiores a 19ºC bastante favoráveis ficaram de acordo com os resultados obtidos por Montoya & Chaves (1974) e por Kushalappa *et al.* (1983). Os primeiros autores indicaram que o ponto mínimo de germinação seria encontrado em temperaturas inferiores a 18ºC e o ponto máximo na temperatura estimada de 23,7ºC. Kushalappa *et al.* (1983) consideraram 14ºC como limite mínimo de atividade do patógeno. A árvore de decisão não identificou efeito negativo de temperaturas acima da ótima no poder germinativo de *H. vastatrix*, como observaram os autores citados. A razão disso pode ser atribuída ao valor máximo de THUR95_PINF (20,3ºC) ter ficado abaixo da temperatura ótima de germinação em todo o período analisado.

Os testes na árvore de decisão baseados na carga pendente de frutos confirmaram a predisposição das plantas à infecção de *H. vastatrix* devido à alta produção (Kushalappa & Eskes, 1989). Segundo Zambolim *et al.* (2002), quanto maior a produção, maiores a incidência e a severidade da ferrugem. As decisões na árvore refletiram as diferenças nos níveis de taxa de infecção observadas entre as lavouras com carga pendente alta e com carga pendente baixa (Figura 2).

O espaçamento entre as plantas é considerado um fator de interferência no progresso da ferrugem do cafeeiro, provavelmente influenciando as condições microclimáticas dentro da lavoura (Kushalappa & Eskes, 1989). Entretanto, o espaçamento nas lavouras de café não foi significativo na determinação da classe da taxa de infecção da ferrugem. A variável LAVOURA não ter aparecido em nenhum teste na árvore de decisão refletiu também o comportamento do progresso da doença (Figura 2), que não exibiu nenhuma distinção evidente devido ao espaçamento.

Temperaturas elevadas no período de incubação exerceram efeito negativo nas taxas de infecção da ferrugem do cafeeiro, efeito esse observado pelos testes em TMAX_PI_PINF e TMIN_PI_PINF na árvore de decisão. Moraes *et al.* (1976) observaram que temperaturas médias máximas micro-climáticas acima de 31°C ocasionaram efeito depressivo sobre o período de incubação. Essas temperaturas corresponderam a temperaturas médias máximas macro-climáticas, obtidas em posto meteorológico, por volta de 28°C, bem próximo dos 28,65°C estabelecidos pela árvore de

TABELA 2 - Medidas de avaliação da árvore de decisão

	Método de estimativa	
Medida	**Resubstituição**[a]	**Validação cruzada**[b]
Taxa de erro	0,22	0,27
Acurácia	0,78	0,73

[a]Estimativa das medidas de avaliação sobre o conjunto de dados de treinamento.
[b]Estimativa das medidas correspondem à média de avaliação sobre 10 partições aleatórias do conjunto de dados de treinamento.

TABELA 3 - Matriz de confusão da árvore de decisão

TAXA_INF3N		Predita[a]			
		TX1(<=0)	**TX2(>0<=5)**	**TX3(>5)**	**TOTAL**
Verdadeira[b]	**TX1(<=0)**	135	13	6	154
	TX2(>0<=5)	16	47	19	82
	TX3(>5)	13	14	101	128
	TOTAL	164	74	126	364

[a]As colunas da tabela representam as classes preditas pelo modelo.
[b]As linhas da tabela representam as classes verdadeiras no conjunto de treinamento.

decisão no teste sobre TMAX_PI_PINF (Figura 4, nó 13).

Quando THUR95_PINF não foi favorável à infecção (menor que 17°C), o limiar que determinou o efeito inibidor da temperatura no período de incubação mostrou-se menor: TMAX_PI_PINF >= 25,95°C (Figura 4, nó 5). Isto parece indicar que germinações ocorridas em condições menos favoráveis são mais sensíveis ao efeito da temperatura no período de incubação. Montoya & Chaves (1974) observaram que temperaturas menos favoráveis à germinação (18 e 26°C) prolongaram o período de incubação (referido como período de geração) e diminuíram o nível de infecção da ferrugem, enquanto temperaturas mais favoráveis à germinação (20, 22 e 24°C) proporcionaram períodos de incubação mais curtos e maior número de ciclos de infecção. Segundo esses autores, as condições que afetaram o processo germinativo, além de terem determinado uma maior ou menor porcentagem de germinação, também exerceram influência na colonização do fungo no tecido vegetal. Sendo assim, de acordo com os resultados do presente trabalho, dependendo das condições de germinação, a colonização do fungo no tecido vegetal sofreu influência diferenciada das condições do ambiente, especificamente da temperatura, acrescentando-se à hipótese de Montoya & Chaves (1974).

A água da chuva é importante para a germinação dos esporos. A chuva, normalmente associada com o vento, é o principal agente de disseminação dos uredósporos (Kushalappa & Eskes, 1989; Zambolim *et al.*, 1997). As chuvas de baixa intensidade e o orvalho que umedecem as folhas durante várias horas são especialmente propícios (Montoya & Chaves, 1974). Por outro lado, chuvas fortes em um curto período de tempo podem conduzir a maior parte dos esporos para o chão (Kushalappa & Eskes, 1989). Esse comportamento irregular pode ter sido a razão das variáveis relacionadas com a precipitação (MED_PRECIP_PINF e PRECIP_PINF) não terem aparecido com grande importância na árvore de decisão. A umidade relativa média diária (UR_PINF) parece ter expressado melhor a importância das chuvas. As estações chuvosas estão frequentemente associadas com alta umidade relativa do ar (Kushalappa & Eskes, 1989). O único teste na árvore de decisão com base em MED_PRECIP_PINF (Figura 4, nó 12) pode ter expressado o efeito mencionado das fortes chuvas. Valores de MED_PRECIP_PINF maiores ou iguais a 2,15 mm (ou PRECIP_PINF >= 73 mm) causaram efeito depressivo nas taxas de infecção da ferrugem do cafeeiro.

Médias baixas de temperatura mínima diária e médias altas de temperatura máxima diária (TMIN_PINF e TMAX_PINF, respectivamente) exerceram influência negativa nas taxas de infecção da ferrugem do cafeeiro. Isso mostrou que não só a temperatura durante o molhamento foliar foi importante; as máximas e mínimas de temperatura nos dias do período de infecção também exibiram importância, embora em grau bem menor.

No Brasil, é freqüente a presença de água livre na superfície das folhas do cafeeiro, mesmo no inverno, estação seca, principalmente devido ao orvalho; as baixas temperaturas enquanto a folha está molhada torna-se o fator limitante para a germinação e a penetração do fungo (Kushalappa & Eskes, 1989). É o que parece ter sido capturado pela árvore de decisão. Os períodos de molhamento foliar prolongado (NHUR95_PINF e NHNUR95_PINF), presentes em praticamente todos os períodos de infecção, não serviram à árvore de decisão para identificar aqueles com maiores ou menores taxas de infecção. A temperatura média nos períodos de molhamento foliar (THUR95_PINF) é que foi escolhida como o fator determinante das classes de taxa de infecção. A temperatura média nos períodos noturnos de molhamento foliar (THNUR95_PINF) foi sempre muito próxima de THUR95_PINF. Por isso, tiveram importância equivalente para o algoritmo de indução: a variável THUR95_PINF chegou a ser escolhida em um dos testes da árvore de decisão apenas por aparecer antes de THNUR95_PINF no conjunto de treinamento.

Na avaliação geral, a árvore de decisão classificou corretamente 283 exemplos de um total de 364 do conjunto de treinamento, ou seja, 78% dos exemplos a partir dos quais a árvore de decisão foi gerada tiveram a classe de taxa de infecção predita igual à classe verdadeira. Por classe de taxa de infecção, o desempenho da árvore para as classes 'TX1(<=0)' e 'TX3(>5)' foi acima da média. Foram classificados corretamente 88% dos exemplos da classe 'TX1(<=0)' e 79% dos exemplos da classe 'TX3(>5)'. O menor desempenho para a classe 'TX2(>0<=5)', com 57% dos exemplos classificados corretamente, talvez esteja relacionado com o menor número de exemplos desta classe no conjunto de treinamento (Figura 3), permitindo que as

outras duas classes prevalecessem na distribuição final dos exemplos nos nós folhas.

O desempenho da árvore de decisão para classificação de outros exemplos, nos quais a árvore não foi treinada, foi estimado em 73% de acurácia. Esse valor de acurácia é uma estimativa mais confiável de desempenho, caso se quisesse avaliar o potencial de uso da árvore de decisão como modelo de alerta da ferrugem do cafeeiro. Cabe ressaltar que um eventual uso da árvore de decisão como modelo de alerta deveria ser restrito à região onde os dados analisados foram obtidos ou a regiões com características parecidas do ambiente estudado.

Concluindo, a técnica de indução de árvores de decisão se mostrou uma ferramenta adequada para o estudo de epidemias da ferrugem do cafeeiro. A árvore de decisão apresentada demonstrou todo o seu potencial como modelo de representação simbólica e interpretável. Permitiu a identificação das fronteiras de decisão presentes nos dados e da lógica contida neles, auxiliando na compreensão de quais variáveis explicativas, e de como as interações dessas variáveis conduziram a interpretar a taxa de progresso da ferrugem no campo.

AGRADECIMENTOS

À Fundação Procafé por ceder os dados relacionados com o monitoramento da incidência da ferrugem do cafeeiro, em especial ao Engenheiro Agrônomo Leonardo Bíscaro Japiassú. Ao SAS Brasil pela concessão da licença de uso do SAS® Enterprise Miner™ por meio de seu Programa Acadêmico.

REFERÊNCIAS BIBLIOGRÁFICAS

Apte C, Weiss S (1997) Data mining with decision trees and decision rules. Future Generation Computer Systems 13:197-210.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Boca Raton FL. CRC Press.

Butt DJ, Royle DJ (1974) Multiple regression analysis in the epidemiology of plant diseases. In: Kranz J (Ed.) Epidemics of plant diseases: mathematical analysis and modeling. New York NY. Springer Verlag. pp. 78-114.

Chalfoun SM (1997) Doenças do cafeeiro: importância, identificação e métodos de controle. Lavras MG. FAEPE, Universidade Federal de Lavras.

Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000) CRISP-DM 1.0: step-by-step data mining guide. Chicago IL. SPSS.

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (Eds.) Advances in knowledge discovery and data mining. Menlo Park CA. AAAI Press. pp. 1-34.

Gleason ML, Taylor SE, Loughin TM, Koehler KJ (1994) Development and validation of an empirical model to estimate the duration of dew periods. Plant Disease 78:1011-1016.

Han J, Kamber M (2001) Data mining: concepts and techniques. San Francisco CA. Morgan Kaufmann.

Hand D, Mannila H, Smyth P (2001) Principles of data mining. Cambridge. MIT Press.

Japiassú LB, Garcia AWR, Miguel AE, Carvalho CHS, Ferreira RA, Padilha L, Matiello JB (2007) Influência da carga pendente, do espaçamento e de fatores climáticos no desenvolvimento da ferrugem do cafeeiro. Anais, 5º. Simpósio de Pesquisa dos Cafés do Brasil, Águas de Lindóia SP. CD-ROM.

Kim KS, Taylor SE, Gleason ML, Koehler KJ (2002) Model to enhance site-specific estimation of leaf wetness duration. Plant Disease 86:179-185.

Kushalappa AC, Akutsu M, Ludwig A (1983) Application of survival ratio monocyclic process of *Hemileia vastatrix* in predicting coffee rust infection rates. Phytopathology 73:96-103.

Kushalappa AC, Eskes AB (1989) Coffee rust: epidemiology, resistance, and management. Boca Raton FL. CRC Press.

Molineros JE, Madden L, Lipps P, Shaner G, Osborne L, Shaukat A, Francl L, de Wolf ED (2004) Comparison of methods for developing fusarium head blight forecasting models. In: Canty SM, Boring T, Wardwell J, Ward RW (Eds.) Proceedings, 2nd International Symposium on Fusarium Head Blight, Orlando FL. p. 475.

Molineros JE, de Wolf ED, Francl L, Madden L, Lipps P (2005) Modeling epidemics of fusarium head blight: trials and tribulations. Phytopathology 95(Suppl.):71.

Monard MC, Baranauskas JA (2002a) Indução de regras e árvores de decisão. In: Rezende SO (Org.) Sistemas inteligentes: fundamentos e aplicações. Barueri SP. Editora Manole. pp. 115-139.

Monard MC, Baranauskas JA (2002b) Conceitos sobre aprendizado de máquina. In: Rezende SO (Org.) Sistemas inteligentes: fundamentos e aplicações. Barueri SP. Editora Manole. pp. 89-114.

Montoya RH, Chaves GM (1974) Influência da temperatura e da luz na germinação, infectividade e período de geração de *Hemileia vastatrix* Berk. & Br. Experientiae 18:239-266.

Moraes SA (1983) A ferrugem do cafeeiro: importância, condições predisponentes, evolução e situação no Brasil. Campinas SP. Instituto Agronômico.

Moraes SA, Sugimori MH, Ribeiro IJA, Ortolani AA, Pedro Junior, MJ (1976) Período de incubação de *Hemileia vastatrix* Berk. et Br. em três regiões do Estado de São Paulo. Summa Phytopathologica 2:32-38.

Paul PA, Munkvold GP (2004) A model-based approach to preplanting risk assessment for gray leaf spot of maize. Phytopathology 94:1350-1357.

Pinto ACS, Pozza EA, Souza PE, Pozza AAA, Talamini V, Boldini JM, Santos FS (2002) Descrição da epidemia da ferrugem do cafeeiro com redes neuronais. Fitopatologia Brasileira 27:517-524.

Quinlan JR (1993) C4.5: programs for machine learning. San Francisco CA. Morgan Kaufmann.

Silva-Acuña R, Zambolim L, Cruz CD, Vale FXR (1998) Estudo epidemiológico da ferrugem do cafeeiro (*Hemileia vastatrix*)

utilizando a análise de trilha. Fitopatologia Brasileira 23:425-430.

Sutton JC, Gillespie TJ, Hildebrand PD (1984) Monitoring weather factors in relation to plant disease. Plant Disease 68:78-84.

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. 2nd Ed. San Francisco CA. Morgan Kaufmann.

Zambolim L, Vale FXR, Pereira AA, Chaves GM (1997) Café (*Coffea arabica* L.): controle de doenças – doenças causadas por fungos, bactérias e vírus. In: Vale FXR, Zambolim L (Eds.) Controle de doenças de plantas: grandes culturas. Vol. 1. Viçosa MG. UFV. pp. 83-140.

Zambolim L, Vale FXR, Costa H, Pereira AA, Chaves GM (2002) Epidemiologia e controle integrado da ferrugem-do-cafeeiro. In: Zambolim L (Ed.) O estado da arte de tecnologias na produção de café. Viçosa MG. UFV. pp. 369-450.

Recebido 2 Outubro 2007 - Aceito 19 Março 2008 - TPP 7115
Editor Associado: Laércio Zambolim